DOCUMENT RESUME

ED 204 399 TM 810 403

AUTHOR Delucchi, Kevin L.
TITLE The Use and Misuse of Chi-square: Lewis and Burke Revisited.
PUB DATE Apr 81
NOTE 54p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Research: *Error of Measurement: *Goodness of Fit: *Literature Reviews: Mathematical Formulas: Maximum Likelihood Statistics: *Research Problems: Statistical Analysis
IDENTIFIERS Burke (C J): *Chi Square Test: *Contingency Tables: Lewis (D): Power (Statistics): Qualitative Data

ABSTRACT

The proper use of Pearson's chi-square for the analysis of contingency tables is reviewed. A 1949 article by Lewis and Burke, in which they cite nine primary sources of error in the use of chi-square, serves as the basis of the review. Those nine sources of error are re-examined in light of current research. In addition, techniques and research on chi-square published after 1949 are discussed. Emphasis is placed on techniques which are of use to the practical researcher in education who often deals with qualitative ordered and unordered data. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED204399

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Use and Misuse of Chi-square: Lewis and Burke Revisited

Kevin L. Delucchi

Department of Education

University of California, Berkeley

April 1981

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. L. Delucchi

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 903

The first paper to describe chi-square was published by Karl Pearson in 1900. As noted by Cochran (1952), the chi-square paper was, and still is, one of the most important publications in the history of modern statistics.

In 1949 Lewis and Burke authored an article appearing in the Psychological Bulletin entitled, "The Use and Misuse of the Chi-Square Test." Their stated aim was to counteract the improper use of this statistic by psychologists. The paper addressed nine major sources of error, cited examples from the literature to illustrate these points, and caused a stir among practicing researchers. Subsequently, the Lewis and Burke paper was followed by several responses (Edwards, 1950; Pastore, 1950; Peters, 1950) and a rejoinder by Lewis and Burke (1950).

Since then, a great deal of research has been conducted on the chi-square procedure and several methods have been developed to handle some of the problems cited by Lewis and Burke. This paper is a review of that literature. It is an attempt to address the problems listed by Lewis and Burke in light of current knowledge and to form recommendations regarding the use and misuse of the chi-square test.

Background

In a compact writing style, Karl Pearson used a geometric proof to derive the distribution theory for establishing the necessary significance level for testing the chi-square statistic. He concerned himself specifically with the problem of determining goodness-of-fit and gave eight numerical illustrations of the use of this new criterion. It is interesting to note that he did not show that the limiting distribution

of the test statistic is χ^2 . This fact was proven subsequently (Cramér, 1946). Pearson also provided incorrect degrees of freedom for testing the statistic.

Originally, the value calculated for a test statistic was compared against tabled values such as those in Elderton's Tables of Goodness-of-Fit (Pearson, 1914). The table was entered using n' = r(c), where r = number of rows and c = number of columns in the contingency table. But in 1915 Greenwood and Yule published an article on research into the effect of inoculation against typhoid and cholera in which they noted that a comparison of proportions should yield the same result as a chi-square test, but it did not. Unable to explain this discrepancy, they stated a preference for the more conservative chi-square procedure. This same inconsistency was noted by Bowley (1920). The determination of the correct degrees of freedom as (r-1)(c-1) was shown by Fisher in two theoretical papers (1922, 1924) and confirmed by Yule (1922) and Brownlee (1924) using sampling experments.

As the use of the chi-square procedure began to grow, its applications and limitations were explored. In the first of three associated papers, Fry (1938) presented and explained the derivation of the chi-square statistic. Subsequent to Fry, Berkson's (1938) paper pointed to the fact that as the sample size increases, the test statistic will eventually reach a significant level. Berkson also noted that this is basically an omnibus test of the hypothesis of equal proportions. That is, one could not locate the specific source within a design that produced a significant result. These two papers were in turn followed by a discussion by Camp (1938) regarding further interpretation of chi-

square.

Many significant contributions to both the theory and applications of this test statistic followed within the next 30 to 40 years. Certainly the most important contributors include Karl Pearson himself, R. A. Fisher, J. Neyman, and E. S. Pearson. A brief historical development can be found in H. O. Lancaster's book along with an excellent bibliography (1969).

The Chi-Square Statistic

Following the lead of Lewis and Burke, this paper is written with the social science researcher in mind. Consequently, the mathematical derivations are more appropriately handled elsewhere (Cramér, 1946; Lancaster, 1969). The following basics of the derivation are presented following Fry (1938). To avoid confusion, the symbol X^2 will be used to distinguish the calculated test statistic from the tabled distribution represented by the Greek symbol χ^2, against which the X^2 value is compared in hypothesis testing.

Given a population of M independent events with s possible outcomes, the joint probability $P(n_j)$ of obtaining n_1 events in category 1, n_2 events in category 2, and so on up to n_s events in category s is given by the multinomial distribution function

$$P(n_j) = \frac{M!}{n_1!\, n_2! \cdots n_s!}\, p_1^{n_1} p_2^{n_2} \cdots p_s^{n_s} \qquad (1)$$

where s is the number of categories or possible outcomes. From this expression the distribution function of chi-square is derived.

This exact formula (1) is very difficult to compute. However, a reasonable approximation may be substituted. This is accomplished by three approximations in the formula. The first involves replacing the factorials by their Stirling approximations. This produces

$$P(n_j) = \frac{1}{\sqrt{(2\pi m)^{s-1}}\sqrt{p_1 p_2 \cdots p_s}} \left(\frac{m p_1}{n_1}\right)^{n_1+\frac{1}{2}} \left(\frac{m p_2}{n_2}\right)^{n_2+\frac{1}{2}} \cdots \left(\frac{m p_s}{n_s}\right)^{n_s+\frac{1}{2}} \quad (2)$$

The second approximation is equivalent to replacing $\left(1 + \frac{x}{m}\right)^m$ by e^x for large m. In this case the result is

$$\left(\frac{m p_j}{n_j}\right)^{n_j+\frac{1}{2}} = e^{-x_j\sqrt{m p_j} - \frac{1}{2}x_j^2} \quad (3)$$

The third approximation consists of replacing the sum of the discrete probabilities of each n_j by an integral. The computational result is well known as

$$X^2 = \sum_{i=1}^{s} \frac{(x_i - e_i)^2}{e_i} \quad (4)$$

where x_i is the observed frequency in class i and e_i is the expected frequency in class i which equals np_i.

Two items which enter into the discussion concerning the proper use of chi-square should be noted at this point. First, to employ the underlying multinomial distribution, the assumption that the x_i are distributed normally is necessary. This means that the expected values (e_i) must be sufficiently large enough for the approximation to be adequate. Second, equation (1) also requires that each of the

probabilities in that expression be independent. This implies that the terms which are summed in equation (4) must be independent of each other.

The Use and Misuse of Chi-Square

Lewis and Burke centered their 1949 article around nine principal sources of error they found in their review of published research. Those nine sources are:

1) Lack of independence among single events or measures
2) Small theoretical frequencies
3) Neglect of frequencies of non-occurrence
4) Failure to equalize the sum of the observed frequencies and the sum of the theoretical frequencies
5) Indeterminant theoretical frequencies
6) Incorrect or questionable categorizing
7) Use of non-frequency data
8) Incorrect determination of the number of degrees of freedom
9) Incorrect computations

This paper will address each of these issues and then consider some aspects of the chi-square procedure that Lewis and Burke did not list as sources of error.

Lack of Independence Among Single Events or Measures

In order for the limiting distribution of X^2 to be χ^2 it is necessary that those events or measures from which X^2 is calculated be independent. This is so because it is the joint probability of n independent events that is given by the multinomial distribution func-

tion.

In designs involving single subject research, or repeated measures on several subjects, this lack of independence is obvious. But often a lack of independence is not noticed, particularly when the final X^2 value is the result of the addition of several other X^2s. A subtle yet telling example is cited by Lewis and Burke early in their paper.

In a hypothetical experiment twelve dice were thrown 14 times and the number of "ones" appearing on each throw were recorded. The test statistic was calculated by summing the quantity $\frac{(X_i - e_i)^2}{e_i}$ for each of the 14 throws. The problem with this procedure is that the same twelve dice were thrown each time. There is no independence between the terms which are summed. Therefore, statements pertaining to any population, other than the 12 dice themselves, cannot be meaningfully made. If one wishes to generalize the results beyond these twelve dice, then a new sample must be drawn. In his response to Lewis and Burke, Peters (1950) makes this point and remarks that a lack of generalizability to a population is probably not too useful to most researchers. But Peters holds firm in stating that if one is concerned with <u>these</u> dice, or subjects, then repeated measures are appropriate.

<u>Small Theoretical Frequencies</u>

One of the most controversial aspects regarding the use of the chi-square procedure is the establishment of a minimum expected value. That is, a value below which the smallest expected frequency may not drop for the application of the test to be appropriate. This is required by the use of the three approximations in the derivation. In

order for a calculated X^2 to approximate χ^2 it is necessary for the sample to be of sufficient size to make those approximations reasonable. This is reflected by the expected value in each cell.

Lewis and Burke called the use of expected frequencies which are too small the most common weakness in the use of chi-square (p. 460). In their paper, they took the position that expected values of five were probably too low. They stated a preference for a minimum expected value of 10 with five as the absolute lowest limit. Lewis and Burke subsequently cited two published studies each employing a chi-square test with expected values below 10 as examples. It appears today that their position, a popular one among researchers, may be overly conservative.

This problem has been examined from two different perspectives. One may consider this issue in relation to the use of chi-square for testing goodness-of-fit. In this approach, as the categories are chosen arbitarily, the researcher has control over the size of the expected value by choice of the category size. In contrast, the categories of contingency tables are relatively limited and one is forced to increase the expected values by increasing the sample(s) size and/or collapsing rows and/or columns. However, it is often difficult, if not impossible, to collect more data to increase N. Collapsing columns and/or rows is in effect throwing away information. Additionally, the information is in an area, the extremes, where differences are most likely to occur. Research taken from the perspective of this later case will be considered first.

Recommendations vary a great deal. Kendall (1952) prefered expected frequencies greater than 20. Cramér (1946) has recommended

values be greater than 10, Fisher (1938) preferred a lowest value of five. Jeffreys (1961), Slakter (1965), and Kempthorne (1966) set one as the minimal expected frequency allowable. Wise (1963) has taken the stance that the expected cell frequencies could be quite small if they are nearly equal to each other. In fact, Wise recommended small but equal expected frequencies over the case where a few expected values are small and the remaining frequencies are well above most criteria.

In a 1952 article Cochran suggested that instead of a single value, the application of chi-square may be deemed appropriate if no more than 20% of the cells have expected values between one and five. Good, Grover, and Mitchell (1970) concluded that in the case where each of the s categories has a probability of 1/s, an equiprobable distribution, the approximation of the test statistic to the chi-square distribution is adequate even when the expected values are as low as 1/3 (p. 275). This apparent robust nature of the procedure is also supported by Lewontin and Felsenstein (1965). They used Monte Carlo methods to examine 2 x N tables with fixed marginals. When the expected values are small in each cell the authors concluded that the test tends to be conservative when the degrees of freedom equal, or exceed, five. Lewontin and Felsenstein found that even the occurrences of expected values below one generally do not invalidate the procedure.

The examination of this problem from the more flexible perspective of the use of chi-square in the case of testing goodness-of-fit has produced some interesting results. Kendall and Stuart (1952) following suggestions by Mann and Wald (1942) and Gumbel (1943) recommended that one choose categories so that each has an expected frequency equal to

the reciprocal of the number of categories. They prefer a minimum value of five. In 1968, Slakter presented the results of a Monte Carlo study concerning the accuracy of an approximation of power for the chi-square goodness-of-fit test with small but equal expected frequencies. He used various combinations of sample size, number of categories, and Type I error probability levels. The results confirmed his earlier work (1965, 1966) and the work of Good (1961) and Wise (1963) which indicated that the nominal alpha level does not deviate substantially when the expected values are small but equal.

In an article based on his dissertation, Yarnold (1970) numerically examined the accuracy of approximation of the chi-square goodness-of-fit. He proposed that, "If the number of classes, s, is three or more, and if r denotes the number of expectations less than five, then the minimum expectation may be as small as 5r/s" (p. 865). The remainder of his paper deals with a new approximation technique used to study the proposed rule. In conclusion, he stated that, "One of the main conclusions of this article is that the upper one and five percentage points of the X^2 approximation can be used with much smaller expectations than previously considered possible" (p. 882).

After considering earlier work, Roscoe and Byars (1971) recommended that for the examination of goodness-of-fit with more that one degree of freedom, one should be concerned with the <u>average</u> expected value. In the uniform case, that is, equal expected cell frequencies, they suggest an average value of two or more for an alpha equal to .05 and four or more for an alpha equal to .01. They exhort the use of this average expected value rule in the test for independence as well, even when the

sample sizes are not equal.

The advantages of several goodness-of-fit tests for discrete data were reviewed by Horn (1977). In so doing she pointed out that Roscoe and Byars rule is in agreement with Slakter's (1965, 1966) suggestion that what may be most important is the average of the expected frequencies. She also noted that this subsumes Cochran's rule that 20% of the expected frequencies should be greater than one.

There is a further point which should be mentioned. As Horn points out, the chi-square goodness-of-fit test is an approximation in two ways. It approximates the exact multinomial goodness-of-fit test and its distribution is an approximation to the theoretical chi-square distribution. The studies cited above are concerned with the second form of approximation. Tate and Hyer (1973) have explored the accuracy of approximation in the first form. They stated that "To justify the X^2 test because the X^2 distribution tails off similarly to the [theoretical] chi-square distribution is to assume that chi-square is itself an accurate approximation to the multinomial" (p. 837).

Tate and Hyer (1969) generated 162 various multinomial distributions and compared then to chi-square values. Their 1973 paper examined their data more closely. They concluded that the chi-square procedure produces false results for a given alpha when expected values drop below 10. They noted that the degree of accuracy required will vary from situation to situation. When close approximation to the exact multinomial is needed, chi-square should only be used when the expected frequencies are above 20 or so.

Most recently, Overall (1980) has examined the effect of low expected frequencies in one row or column of a 2 x 2 design on the power of the chi-square statistic. This most often results from the analysis of infrequently occurring events. Setting $1 - \beta = .70$ as a minimally acceptable level, Overall concluded that when expected values are quite low, the power of the chi-square test drops to a level that produces a statistic which, in his view, is almost useless. Further considerations regarding the power of chi-square may be found in Cramér (1946), Bennett and Hsu (1960), Harkness and Katz (1964), Chapman and Meng (1966), and Broffitt and Randles (1977).

As a general rule it seems that the chi-square statistic may be properly used in cases where the expected values are much lower than previously considered permissible, although this is not always true as Tate and Hyer and Overall have shown. The practitioner must take into consideration the level of precision required by his work. The closer one desires to be to the exact probabilities of the multinomial, the larger the sample sizes and expected values must be. For most applications, Cochran's rule which states that all expected values be greater than one, and not more than 20% be less than five, offers a fair balance between practicality and precision. The more exploratory the research, the more one may relax this rule. It also seems appropriate to relax this rule if the expected values, though small, are roughly equal.

Neglect of Frequencies of Non-Occurrence and Failure to Equalize the Sum of Observed and Expected Values

In his reply to Lewis and Burke, Peters (1950) took exception to the propriety of claiming that these aspects are sources of error.

Peters stated that one's research questions determined whether the frequencies of non-occurrence should be included in the calculations. He further stated that it is the computational formula

$$X^2 = \sum_{i=1}^{s} \frac{(x_i - e_i)^2}{e_i} \qquad (4)$$

which requires that the sum of observed and expected frequencies be equal, but not the generalized definition which includes the true population mean and variance.

In their rejoinder, Lewis and Burke (1950) show why they were correct in listing both of these points as errors. The basis is a proof of an underlying theorem of chi-square shown by Cramér (1946) and its generalization. Given the asumption that the frequencies for all possible outcomes are used and that the sum of the observed frequencies equals the sum of the theoretical, Cramér's proof hold for equation (4).

Therefore, in all applications of this formula, the sum of the observed frequencies must equal the sum of the expected. In addition, frequencies for all of the possible outcomes must be included in the calculation. That is, in a test of the homogeneity of several groups based on the number within each group having a certain property, the calculation of chi-square must include the frequencies of those sample members, within each group, who do not have that property.

A recent example of this was cited by Slaughter and Marascuilo (Note 2). In 1979 Scheuneman presented a method for assessing bias in test items using a modified chi-square procedure. Basically, the procedure involves dividing the number of correct responses to an item, by

group, into several categories based on total raw score. A chi-square is then calculated to test that the proportions passing an item, within the ability categories, are the same for each group. An item is defined as biased if the chi-square value is significant.

Slaughter and Marascuilo point out, for the statistic to approximate the chi-square distribution, its calculation must include the frequency in each group that failed the item. Scheuneman makes reference to this when she states,

> "It should be noted, however, that because the modified procedure does not include incorrect responses, the obtained distribution of chi-square values may not always approximate the chi-square distribution, particularly if the sample sizes for the groups being compared are quite different, or the cell frequencies are very large" (p. 147).

But she does not indicate why she does not use the more exact procedure.

Slaughter and Marascuilo demonstrate the proper use of this procedure and indicate that a substantial number of the items judged as fair, by her definition, are in fact biased. Given that this method of assessing item bias is itself somewhat rough, there is no justification for weakening it even further by excluding the incorrect responses as Scheuneman proposes.

Indeterminant Theoretical Frequencies

It is possible that the theoretical frequencies, the expected values against which each observed value is compared, may not be

calculable. Lewis and Burke have illustrated such a case in a hypothetical coin guessing experiment in which subjects recorded their guesses as to whether a head or tail would appear on each of four tosses. The number of correct guesses, ranging from zero to four were compared to the expected frequencies generated by the binomial distribution function. Since the four guesses of each subject could not be considered independent of one another, the theoretical distribution is clearly not binomial. In this case the most one could do would be to test the obtained distribution values against values given by some other research using the same experimental design.

As can be seen, this problem of indeterminant theoretical frequencies arises in the test for goodness-of-fit where category choice is arbritrary. In their final paragraph on this subject, Lewis and Burke offer a guideline for deciding if the theoretical frequencies are indeed calculable. They state,

> "It is usually true that theoretical frequencies are incalculable if the observed frequencies are in any way related, and also if mutually contradictory assumptions can be made, with about equal justification, concerning the likelihood of occurrence or non-occurrence the events (resonses) that yielded the observed frequencies" (p.483).

Incorrect or Questionable Categorizing

In deciding upon the categories to be used, care must be taken in their selection - especially when the choice is arbitary. The

value of the test statistic will be unduly inflated if one or more of the categories contains a substantial number of obervations in only one cell of that category. Lewis and Burke provide an excellent example of this.

In a study comparing the drawings of normal and abnormal subjects one of the categories for classifying the drawings was labeled, "fantastic compositions". As one would expect, all 26 of the drawings placed in this class were drawn by abnormal subjects. The individual X^2 value for this group (26.0) accounted for 25% of the total X^2 (99.6) even though only 5% of the total frequencies fell into this category. Lewis and Burke offer two general rules to follow which should help in dealing with this problem: 1) categories for frequency data should be established, whenever possible, on the basis of completely external criteria, and 2) information on the reliability of the categories should be offered. This becomes very important as the choice of categories becomes more and more arbritary. A careful, logical examination of a study design, such as the one mentioned above, may not always be possible.

Use of Non-Frequency Data

A simple example will show that the formula,

$$X^2 = \sum_{i=1}^{s} \frac{(x_i - e_i)^2}{e_i} \qquad (4)$$

can only be applied to frequency data. Given an observed frequency of four and an expected frequency of two we have for the single

term

$$\frac{(x-e)^2}{e} = \frac{(4-2)^2}{2} = 2$$

Let us assume that the four and two are measures on some scale such as pounds, inches, or even a ratio such as errors per minute. If one were to change the scale of measurement by converting pounds to ounces, inches to feet, or errors per minute to errors per 90 seconds, the value of all terms calculated would change by the same factor. Thus, for example, to double the number of units in the scale of measurement would change the observed value of this hypothetical example from four to eight, the expected value from two to four and the resulting single term would equal

$$\frac{(x-e)^2}{e} = \frac{(8-4)^2}{4} = 4$$

The chi-square value would be doubled soley by this change in metric.

It must be made clear that this is not to say that either the chi-square statistic or the function if its limiting distribution are derived from, or refer only to, frequencies. However, the computing formula (4) can only properly be applied to frequencies of independent observations.

Incorrect Determination of the Number of Degrees of Freedom

One way to interpret the number of degrees of freedom associ-

ated with a contingency table is to note that it represents the number of independent pieces of information contained in the sample about the truth of an hypothesis under test. That is, if we have a set of N numbers which may take on any values with the restriction that they add to a given value, then N-1 of them are free to vary. The one remaining value is determined as it must be that single value which, when added to the sum of the N-1 numbers, equals the value given by the restriction. Thus, N data points with a single restriction have N-1 degrees of freedom. Every restriction imposed decreases the available information contained in the data.

For a contingency table with r rows and c columns, the degrees of fredom equal (r-1)(c-1). This holds regardless of whether one is testing two variables measured on a single group for independence, or whether one has C groups which are being tested for homogeneity across R rows or categories. But this is true for different reasons. Marascuilo and McSweeny (1977) present a discussion of this and the following is taken from their presentation.

In the test of homogeniety, one has an r x c contingency table where the number of columns, c, corresponds to the number of independent samples. As the expected frequencies of the r categories for sample c must add to $n_{.c}$, there are (r-1) degrees of freedom in that one sample. For the c samples, there exist c(r-1) degrees of freedom. In addition, the r proportions are unknown and must be estimated. As they must sum to unity, (r-1) of them are free to vary. The degrees of freedom for the entire table, therefore, equal c(r-1)-(r-1) = (r-1)(c-1).

In the test for independence, only a single sample size is known. The frequencies must sum to this value leaving (rc-1) degrees of freedom. In the case of two variables, the sum of their probabilities must add to one. For r levels of one variable, (r-1) are free to vary. For c levels of the second variable, (c-1) need to be estimated. The entire table thus has

degrees of freedom = (rc-1)-(r-1)-(c-1)

= (r-1)(c-1)

Incorrect Computations

Mechanical errors aside, any of the aforesaid errors would lead, in effect, to an incorrectly computed test statistic. Lewis and Burke noted one computational error in particular that is easy to make and should be guarded against. This error involves the failure to weight by n when proportions are used instead of frequencies.

As mentioned previously, a chi-square value calculated on non-frequency data can be altered by a change in scale. Given the same data a change from meters to centimeters will increase the value of chi-square by a factor of 100. As a proportion is the ratio of observed frequency to total, a chi-square calculated on proportions will be altered by changing the scale. A change of errors per minute to errors per 120 seconds will double the value of chi-square.

Most proportions encountered will be of the form

$$p = \frac{n_{rc}}{n_{.c}}$$

where n_{rc} is the frequency in the cell defined by row r and column c and where $n_{.c}$ is the total frequency for column c. To convert a proportion to a frequency merely requires that the proportion $\frac{n_{rc}}{n_{.c}}$ be weighted by $n_{.c}$. While contingency tables containing proportions are often more interpretable, a chi-square must be calculated using the frequencies from which the proportions were determined.

Additional Issues

Further research regarding the properties of chi-square have been conducted since the publication of the Lewis and Burke paper. Methods have been developed to strengthen the chi-square test. Also, closer examination of its properties, such as the use of a correction for continuity, have been conducted. Perhaps one of the best papers on this subject was written by Cochran (1954). He presented methods for dealing with some specific contingency table designs and probability distributions. In addition to the previously mentioned recommendations regarding minimum expected values, he discussed testing goodnes-of-fit in different distributions, degrees of freedom in 2 x N tables, and combining 2 x 2 tables. The remainder of this paper deals with further issues in the use of chi-square.

Partitioning

At about the same time that Lewis and Burke were writing, the first extensive work on the partitioning of an I x J contingency table into components was being conducted, by Lancaster (1949, 1950). He demonstrated that a general term of a multinomial can be

reduced to a series of binomial terms, each with one degree of freedom. Irwin (1949) presented a formula for exact partitioning which was simplified algebraically by Kimbal (1954) for easier computation. In 1960, Kastenbaum generalized the partitioning procedure to handle cases where some of the desired partitions contained more than one degree of freedom. Castellan (1965) reviewed these partitioning procedures and agrued for their use in place of constructing a series of 2 x 2 tables based on the following two points.

First, in setting up the full contingency table, it is assumed that the marginal totals represent the population values. It is more likely that the marginals for any 2 x 2 table, taken from the full table, will not adequately reflect those population values. Instead, they will reflect a population different from other populations generated from the same table. There will be as many populations represented as there are 2 x 2 tables produced.

Second, following the procedure Castellan presented, the 2 x 2 tables are additive. The sum of their individual chi-square values equals the chi-square value for the original table. This independence of tables produces uncorrelated chi-squares and thus allows for more meaningful interpretation.

Bresnahan and Shapiro (1966) examined methods for partitioning, including the methods for determining possible partitions. They concluded that all forms of a partitioning follow three basic rules: 1) each cell appears alone once and only once, 2) the same combination of cells appear only once, and 3) the dividing lines of

a partition do not hold for other partitions. Following these rules, additional partitioning schemes may be employed. They derive a general equation for the chi-square which may be applied to any table that may result from partitioning. The equation for an I x J table is written as follows:

$$X^2_{(l-1)(m-1)} = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{n_{ij}^2}{e_{ij}} - \sum_{i=1}^{l} \frac{o_{i\cdot}^2}{e_{i\cdot}} - \sum_{j=1}^{m} \frac{o_{\cdot j}^2}{e_{ij}} + \frac{O}{E} \qquad (5)$$

where,

l = the number of rows in the partitioned table

m = the number of columns in the partitioned table

e_{ij} = the expected value for cell ij calculated from the original table

$e_{i\cdot} = \sum_{j=1}^{m} e_{ij}$

$e_{\cdot j} = \sum_{i=1}^{l} e_{ij}$

$o_{i\cdot} = \sum_{j=1}^{m} n_{ij}$

$o_{\cdot j} = \sum_{i=1}^{l} n_{ij}$

$O = \sum_{i=1}^{l} \sum_{j=1}^{m} n_{ij}$

$E = \sum_{i=1}^{l} \sum_{j=1}^{m} e_{ij}$

n_{ij} = the observed frequency in cell ij

Bresnahan and Shapiro advocated the use of this formula in cases where some cells have low expected values. Instead of pooling data or discarding it to raise the low expected values, one can calculate a chi-square based on the table configuration with adequate expected values. This value will be the contribution of that part of the table to the chi-square for the entire table.

Schaffer (1973) has taken exception to the use of these methods of partitioning, claiming that they do not actually test the questions of interest. For example, a 2 x 4 table may be partitioned into three separate tests, each with one degree of freedom. Schaffer then demonstrated that to test the first of the three resulting hypotheses actually entails testing that all three partitions do not contain significant differences against the alternate hypothesis that the first partition is significant and that the other two are not. This results from the fact that the data from the entire table enter the calculation for a portion of the table in the determination of the expected values. She therefore contends, contrary to Castellan, the data from the entire table should not enter into a partition, since the test produced is not the statistic desired.

On the basis of this argument Schaffer proposes the use of the likelihood ratio statistic. Though it does not partition exactly, its use overcomes the problem of testing "inappropriate" hypotheses. Schaffer notes that while there is no evidence for the superiority of one method over another, Pearson's method has historical priority and a greater ease of computation.

Regardless of which method one uses, partitioning increases the amount of information one is able to glean from the data. If the partitions are orthogonal to one another, the information rendered from each partition does not overlap with any other. However, Schaffer's paper presents an interesting quandary.

If one requires a test of a partition, independent of the

structure of the rest of the partitions, then one must use the log-likelihood ratio as she proposed. The lack of additivity of the likelihood ratio may not always be problematic. Often, only one partition is meaningful and/or accounts for much of the total X^2

In such cases, the choice between the use of the log-likelihood statistic and chi-square rests on the alternate hypothesis that is of interest. If one wishes to test a single partition for homogeneity against the hypothesis that it is not homogeneous and the rest of the partitions are, then chi-square is appropriate. If the test is to be done completely independent of the structure of the rest of the table, then the log-likelihood ratio is the method of choice. The log-likelihood ratio has been proposed for use in more than the analysis of partitions as will be discussed in the next section.

Log Likelihood Ratio

An alternative procedure to calculating X^2 to test a hypothesis concerning a multinomial is the use of the likelihood ratio statistic. It is a maximum likelihood estimate labeled G^2 and defined as,

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij} \log_e \left[\frac{X_{ij}}{e_{ij}} \right] \qquad (6)$$

In their test on discrete multivariant analysis, Bishop, Fineberg, and Holland (1975) used log-linear models, as opposed to additive, models for contingency table analysis. As a summary statistic they stated a preference for Maximum Likelihood

Estimators (MLEs) on theoretical grounds. Additionally, practical reasons for the use of this procedure were given:

1. Ease of computation for linear models.
2. MLEs satisfy certain marginal constraints they call intuitive.
3. "The method of maximum likelihood can be applied directly to multinomial data with several observed cell values of zero, and almost always produces non-zero estimates for such cells (an extremely valuable property in small samples)" (p. 58).

They further state,

> "MLEs necessarily give minimum values of G^2 , it is appropriate to use G^2 as a summary statistic...although the reader will observe that, in those samples where we compute both X^2 and G^2 , the difference in numerical value of the two is seldom large enough to be of practical importance" (p. 126).

There are cases where the likelihood ratio statistic may be preferred over chi-square. Such may occur when some expected values are quite small or where the contingency table contains a structural zero. This occurs when a design contains a cell which can never logically, be filled. Bishop, Fineberg, and Holland offer the example of a classification of type of surgery by sex. The cell defined by male-hysterectomy would never contain an entry.

Several investigators have compared X^2 and G^2 in a variety of research situations. Chapman (1976) provides an overview of much

of this research, including the work of Neyman and Pearson (1931), Cochran (1936), Fisher (1950), Good, Grover, and Mitchell (1970), and West and Kempthorne (1972). From these comparisons, neither of the two procedures emerges a clear favorite. When one method is better in some respect than the other, it seems to result from a particular configuration of sample size, number of categories, expected values, and the alternative hypothesis. If a general statement were to be made, it would appear that the log-likelihood ratio statistic tends to produce closer approximation to the χ^2 distribution in many cases. But this statement must be regarded with two considerations in mind.

As most studies on this matter are confined to examining a few of the many possible cross-classification designs where these two statistics might be used, such a statement must be deemed tentative. In some situations neither measure is preferred over the other. In other cases a slight modification in design or sample size may equalize the performance of both statistics. As a result it is very difficult to synthesize this collection of work in order to reach a definitive recommendation valid for all research, or even a majority.

Also, the actual diffferences observed may be so small that they are inconsequential to the researcher. As in the debate over the matter of expected values, before a decision can be made one must place the question within the context of actual practice. The more one's research demands precision, the more closely one should consider any differences in the statistics one may employ.

Further, one should look closest at the research using conditions most similar to one's own design.

Correction for Continuity

In a single paragraph, Lewis and Burke present the correction for continuity noting that it is justified only in the case of a 2 x 2 table. Their treatment of the subject has the air of a proven method which is utilized without question. But questions have arisen since Lewis and Burke regarding the appropriateness of its use.

Since categorical variables are discrete and the chi-square distribution is continuous, a compensation can be made by adding or subtracting 1/2 to each observed frequency so as to move the observed value closer to the expected value. Thus it becomes more difficult to reject the hypothesis under test. Symbolically, the corrected chi-square is written as,

$$X_c^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left[\left(X_{ij} + \frac{1}{2}\right) - \mathcal{E}(X_{ij})\right]^2}{\mathcal{E}(X_{ij})} \qquad (7)$$

In the case of the 2 x 2 table where

$$X^2 = \frac{N\left(X_{11}X_{22} - X_{12}X_{21}\right)^2}{X_{1\cdot}\, X_{\cdot 1}\, X_{2\cdot}\, X_{\cdot 2}} \qquad (8)$$

the correction proposed by Yates (1934) is calculated as

$$X_c^2 = \frac{N\left(\left|X_{11}X_{22} - X_{12}X_{21}\right| - \frac{N}{2}\right)^2}{X_{1\cdot} X_{\cdot 1} X_{2\cdot} X_{\cdot 2}} \qquad (9)$$

The analytical derivation of the correction expressed in (9) is given by Cox (1970).

The disagreement over the use of this correction is based not on its theoretical grounding but on its applicability. Plackett (1964), confirming empirical results of Pearson (1947), argued that the correction is inappropriate if the data come from independent binomial samples. Grizzle (1967) extended Plackett's results to the general case concluding that the correction is so conservative it is rendered useless for practical purposes.

Supporting the use of the corection Mantel and Greenhouse (1968) have taken exception to the views of Plackett and others. They base their objection on two points. First, they state that the proper model for a 2 x 2 table is a fixed marginal total model. In such a model the correction is not overly conservative. Second, the correction improves the probability estimates except in extreme cases. Such cases occur when the hypergeometric (or binomial) distribution deviates from symmetry beyond some fairly extreme level.

Pirie and Hamdan (1972) attempted to stradle the controversy by deriving corrections for continuity for unconditional, that is random marginal, models. For the 2 x 2 test of independence they arrive at a correction of 1/2 instead of N/2 written as

$$\chi^2_C = \frac{N\left(|X_{11}X_{22} - X_{12}X_{21}| - \frac{1}{2}\right)^2}{X_{1.}X_{.1}X_{2.}X_{.2}} \qquad (10)$$

The probability levels resulting from the use of this correction fall between those produced by the uncorrected statistic (8) and

the corrected (9).

This issue was next addressed by Conover (1974a) and several short comments that immediately followed his article. Elaborating on a stance he had taken in 1971, Conover proposed that the correction for continuity should only be used in 2 x 2 tables if the row and column totals are non-random and either one or the other pair of the row or column totals are equivalent to each other. If this is not the case, Conover maintains that the correction is overly conservative. In his response, Mantel (1974) agreed that a fixed-marginal model is appropriate and proposed a separate correction for each tail of the distribution. Conover (1974b) concurred with this method and recommended it be used in place of the Yates correction when the table totals are non-random.

In the subsequent papers Mittinen (1974) agreed with Conover`s position. Starmer, Grizzle, and Sen (1974) presented simulation results which support the contention that when the column totals are non-random, an uncorrected chi-square is to be preferred over the more conservative corrected procedure.

More recently, Everitt (1977) recommended the use of the correction but offered no support for his decision. Camilli and Hopkins (1978), on the other hand, have presented results from a Monte Carlo study confirming the stance taken by Conover, et al. Their results demonstrated that a Yates correction decreases the accuracy of a probability statement when either, or both, of the margins are not fixed.

The consensus seems to be that the correction for continuity becomes overly conservative when either or both of the marginals in a table are random. As this is often the case in social science research, it would appear that the use of the correction should not be given the blanket recommendation that often accompanies it. If strong conservatism is desired and/or the marginal totals in the contingency table being analyzed are fixed values, then the Yates correction should be applied. However, in all other cases one must be cautious in its use as the correction for continuity will produce very conservative probability estimates. When a correction is desired and the table being analyzed does not have fixed marginal values, the work of Pirie and Hamdan should be considered carefully.

Comparison of Two Independent Chi-Squares

Situations may occur in which one may want to test the equality of two independent chi-square values. Knepp and Entwisle (1969) have presented, in tabular form, the one and five percent critical values for this comparison for $\nu = 1$ to 100. They also mention a normal approximation calculated as

$$Z = \frac{\frac{1}{2} X_1^2 - \frac{1}{2} X_2^2}{\sqrt{\nu}} \qquad (11)$$

where X_1^2 and X_2^2 are two independent sample chi-square values, each with ν degrees of freedom. The statistic z is approximately distributed as a unit normal variable.

D'Agostino and Rosman (1971) have offered another simple

normal approximation for comparing two chi-square values in the form of

$$Z = \frac{\sqrt{X_1^2} - \sqrt{X_2^2}}{\sqrt{1 - \frac{1}{4\nu}}} \qquad (12)$$

This approximation was tested by Monte Carlo methods and found to be quite good for cases with $\nu > 2$. For $\nu = 1$ the researcher must use Knepp and Entwistle's tabled values of 2.19 for $\alpha = .05$ and 3.66 for $\alpha = .01$. D'Agostino and Rosman also note that for $\nu > 20$ the denominator in (11) makes little difference and

$$Z = \sqrt{X_1^2} - \sqrt{X_2^2} \qquad (13)$$

may be used in place of (11).

Comparison of Individual Proportions

The chi-square procedure, as Berkson noted in 1938, is an omnibus test. In the case of a test for homogeneity among K groups classified by J levels of the dependent variable A, the hypothesis under test is that

$$H_0: \begin{bmatrix} P(A_1|G_1) \\ P(A_2|G_1) \\ \vdots \\ P(A_J|G_1) \end{bmatrix} = \begin{bmatrix} P(A_1|G_2) \\ P(A_2|G_2) \\ \\ P(A_J|G_2) \end{bmatrix} = \cdots = \begin{bmatrix} P(A_1|G_K) \\ P(A_2|G_K) \\ \\ P(A_J|G_K) \end{bmatrix} = \begin{bmatrix} P(A_1) \\ P(A_2) \\ \\ P(A_J) \end{bmatrix}$$

against the alternative that H is false. If the hypothesis is rejected, one would like to be able to find the contrasts among the proportions that are significantly different from zero.

This may be accomplished by a well known procedure which

allows one to construct simultaneous confidence intervals for all contrasts of the proportions in the design, across groups, while maintaining the specified Type I error probability. The method is an extension of Scheffe's (1959) theorem which is used for the construction of contrasts in the analysis of variance.

If a linear contrast in the population proportions in a contingency table is denoted as ψ, the sample estimate is $\hat{\psi}$ and is defined as

$$\hat{\psi} = \sum_{k=1}^{K} a_k \hat{p}_k \qquad (14)$$

where $\hat{p}_k$ is the proportion in group k and $\Sigma a_k = 0$. The limiting probability is $(1 - \alpha)$ that, for all contrasts,

$$\hat{\psi} - SE_{\hat{\psi}} \sqrt{\chi^2_{K-1:1-\alpha}} < \psi < \hat{\psi} + SE_{\hat{\psi}} \sqrt{\chi^2_{K-1:1-\alpha}} \qquad (15)$$

where

$$SE_{\hat{\psi}} = \sqrt{\sum_{k=1}^{K} a_k \left(\frac{p_k q_k}{n_k}\right)}, \qquad q_k = 1 - p_k \qquad (16)$$

and $\sqrt{\chi^2_{k-1:1-\alpha}}$ is the $(1 - \alpha)^{th}$ percent value from the chi-square distribution at K - 1 degrees of fredom. Some of the earlier work with this procedure may be found in Gart (1962), Gold (1963), and Goodman (1964).

The only drawback to this post hoc procedure is its lack of power relative to a planned set of contrasts. In place of the use

of X^2 followed by post hoc exploration using the confidence interval defined above one may employ a series of planned contrasts. A more powerful procedure results from the use of a Bonferroni type critical value where the Type I error probability is spread over just the contrasts of interest. Such a value may be found in the table given by Dunn (1961). The value $\sqrt{\chi^2_{k-1:1-\alpha}}$ in the confidence interval is replaced by the value taken from Dunn`s table at Q= the number of planned contrasts and $\nu=\infty$.

Analysis of Ordered Categories

In spite of its usefullness, there are conditions under which the use of Pearson's chi-square, although appropriate, is not the optimum procedure. Such a situation occurs when the categories forming a table have a natural ordering. The value of the statistic expressed in (4) will not be altered if the rows and/or columns in a table are permuted. However, if ordering of the rows or columns exists, their order cannot meaningfully be changed. This is information which chi-square is not sensitive to. Instead, the researcher may choose among three alternatives.

If both rows and columns contain a natural ordering, two methods are available. The first is a procedure taken from Maxwell (1961) as modified by Marascuilo and McSweeny (1977). It is used to test for a linear trend in the responses across categories.

The first step is to quantify the categories using any arbritrary numbering system. As the method is independent of the numbers chosen, both Maxwell and Marascuilo and McSweeny recommend

numbers which simplfy the calculations such as the linear coefficients in a table of orthogonal polynomials. These coefficients are then applied to the marginal frequencies to produce the sums and sums of squares for use in calculating a slope coefficient by the usual formula,

$$\hat{\beta} = \frac{N\left(\sum y_i y_i\right) - \left(\sum y_i\right)\left(\sum y_i\right)}{N\sum y_i^2 - \sum y_i^2} \qquad (17)$$

Under the assumption that $\beta=0$, the standard error of $\hat{\beta}$ is calculated as

$$SE_{\hat{\beta}} = \frac{S_{y_i}^2}{(N-1)S_{y_i}^2} \qquad (18)$$

Then the hypothesis of no linear trend may be tested by

$$X^2 = \frac{\hat{\beta}^2}{SE_{\hat{\beta}}^2} \sim X^2_{v-1} \qquad (19)$$

A decomposition of the total chi-square for the contingency table is obtained by taking X^2 (total) - X^2 (due to linearity)= X^2 (residual). This may often be a more meaningul analysis.

A second procedure involves the use of Kendall's (1970) rank tau, corrected for ties. If the observed tau is statistically significant, the hypothesis of no association is rejected. In addition, the statistic itself is a measure of association or array of the data. Further comments are contained in the section of measures of association. When one of the two variables defining a table are ordered, Kruskal and Wallis' (1952) non-parametric one-

way analysis-of-variance procedure may be utilized to test for equality of distributions.

Consider the case of an I x K contingency table where the dimension I is defined by mutually exclusive, ordered categories. The Kruskal-Wallis statistic is based on a simultaneous comparison of the sum of the ranks for the K groups. To apply the statisitc in the case of an I x K table the frequencies within a category along dimension I are considered to be tied an therefore are all assigned a midrank value. One then sums the ranks acorss I, within group k, to obtain the summed ranks used in calculating the statistic.

Measures of Association

As a final, and important note, a few words must be said about measures of association. It needs to be remembered that the value of a chi-square statistic is a function of sample size. To double the size of a sample, barring sample-to-sample fluctuations, will double the size of the associated chi-square. To compensate for this, the data analyst should always calculate an appropriate measure of association. To report probability levels alone is equivalent to reporting the sample size as an indication of the results. A proper measure of association should be included so as to allow for judging the practical, that is the meaningful significance of the findings. While a proper treatment of this topic deserves a paper unto itself, because of the importance of this subject, an outline of the main measures will be included here.

We begin with the general case of an I x J contingency table. If the data are generated from a single sample, then the proper test is one of independence and a measure of association is the mean square contingency coefficient. Designated as ϕ^2, it's sample estimate is calculated as

$$\hat{\phi}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{X_{ij}^2}{X_{i\cdot} X_{\cdot j}} - 1 \qquad (20)$$

As the maximum value $\hat{\phi}^2$ may obtain, $\hat{\phi}^2_{max}$, is the minimum of I-1 or J-1, a correction for this is

$$\hat{\phi}'^2 = \frac{\hat{\phi}^2}{\hat{\phi}^2_{max}} \qquad (21)$$

and is referred to as Cramer's measure of association (Cramer, 1946).

In the cae of a table generated from K-samples, the proper measure of association is given by the work of Light and Margolin (1971). It is a ratio of the sum of squares between the K groups over the total sum of squares. Their measure, R^2_{LM}, is tested for significance by a chi-square statistic calculated as $X^2_{LM}=(N-1)(I-1)R^2_{LM}$ which is tested at $\nu=(I-1)(K-1)$ degrees of freedom. Light and Margolin have shown that their statistic tends to be larger, and therefore more powerful than the ordinary chi-square in the analysis of a K-group design.

When the frequencies of the k-groups are cross-classified by a dependent variable which is ordered, a more appropriate measure of

association has recently been proposed. As noted earlier, this model is analysed by a non-parametric one-way ANOVA. Carr, Marascuilo, and Serlin (Note 1) have proposed a measure which is the ratio of the calculated test statistic to the maximum the statistic can reach. Their measure ranges from zero to unity and it is interpreted just as eta squared is in the parametric ANOVA.

If both variables are ordered, one is presented with a variety of choices beginning with the standard product-moment correlation coefficient. The use of this method is discussed by Kendall and Stuart (1969) and basically involves the assignment of a set of scores to each category. These pre-assigned scores may be just the natural numbers, 1, 2, 3, ..., normal scores, or a normalized score using relative frequencies of the margins as cutting points for assigning values from the normal distribution. The chief disadvantage of this method centers around the fact that the scores are assigned arbitarily and the measure calculated will vary with the scoring system chosen.

The most appealing measure in this case may well be Kendall's measure of disarray, tau (Kendall, 1970). Its use in ordered contingency tables is illustrated by Kendall in his third chapter. Because those data in the same row or column of a table are considered as neither concordant nor discordant in relation to each other, but as tied, tau corrected for ties, τ_c, must be used. A competitor to tau has been proposed by Goodman and Kruskal in the first of their three extensive papers on measures of association in cross classification (Goodman and Kruskal, 1954; 1959; 1963). The

measure, γ , is the same as Kendall`s tau in the numerator. The denominator is the same except in that it excludes tied values. This means that in all cases tau < gamma. The use of tau is recommended becasue the inclusion of the tied data is a more conservative method and tau approaches the normal distribution faster than Spearman`s rank order correlation (Kendall, 1970).

In the case of a 2 x 2 table, the well known measure of association based on chi-square is phi and is calculated as

$$\phi^2 = \frac{X^2}{N} \qquad (22)$$

If Kendall`s tau is calculated for the same table, then it will be seen that phi = tau. An alternative to the use of phi is to employ the odds ratio.

For a 2 x 2 table the categories defining the table may be labeled as A, $\bar{A}$, B, and $\bar{B}$. The probability of observing B, given the pressence of A, can be expressed as

$$\frac{P(B|A)}{P(\bar{B}|A)} \qquad (23)$$

Alternately, the probability of observing B, given the absence of A, is

$$\frac{P(B|\bar{A})}{P(\bar{B}|\bar{A})} \qquad (24)$$

A simple measure of association, apparently first proposed by Corn-

field (1951), is the ratio of these two odds. In the sample the measure is calculated as

$$\hat{\gamma} = \frac{n_{11}\, n_{12}}{n_{12}\, n_{21}} \qquad (25)$$

with a standard error estimated as

$$SE_{\hat{\gamma}} = \hat{\gamma}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}} \qquad (26)$$

A useful discussion of this measure including additional references may be found in Fleiss (1973). The choice between the two coefficients, tau and phi, for the 2 x 2 table is not clear cut and the reader is referred to Fleiss for further discussion.

Summary

At 80 years of age, Karl Pearson`s chi-square statistic remains one of the most useful, versatile, and popular measures for data analysis. Lewis and Burke are two among many authors who have considered its properties and applications and this paper has, hopefully, served as a general review of that literature. In closing, it is interesting to note a couple of points regarding both the misue and use of chi-square.

In spite of the age of the Lewis and Burke article it is unfortunate to discover that many of the errors outlined in their work can be found in research today. Perhaps because the measure is so well known and so easily used, it is also easily misused.

Care must be taken to ensure that when one selects a method to analyse a set of data, one employs the method(s) used correctly. This applies not only to Pearson`s chi-square, but also to evert method used for inferential purposes.

As a final point, it is important to remember that, as noted earlier, several aspects of the chi-square procedures are still subject to debate such as the minimum expected frequencies allowable and the best way to partition a contingency table. Very few things in life are written in granite and the "right" way to analyse a given set of dat is not one of those things. The wise researcher will keep a track of the relevant literature, seek advise from collegaues, and will forsakethe automatic and mechanical application of statistical methods.

Reference Notes

1. Carr, J., Marascuilo, L. A., & Serlin, R. A measure of association for rank tests based on the Kruskal-Wallis model. Department of Education, U. C., Berkeley, Berkeley, California. 94720.

2. Slaughter, R. S., & Marascuilo, L. A. Nonparametric procedures for analyzing test bias: An alternative to Scheuneman`s approach. Department of Education, U. C., Berkeley, Berkeley, California. 94720.

References

Bennett, B. M., & Hsu, P. On the power function for the exact test of the 2 x 2 contingency table. Biometrika, 1960, 47, 393-398.

Berkson, J. Some difficulties in interpretation of the chi-square test. Journal of the American Statistical Association, 1938, 33, 526-536.

Bishop, Y. M. M., Fineberg, S. E., & Holland, P. W. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press, 1975.

Bowley, A. L. Elements of Statistics. London: P. S. Kings & Sons, 1920.

Bresnahan, J. L., & Shapiro, M. M. A general equation and technique for the exact partitioning of chi-square contingency tables. Psychological Bulletin, 1966, 66, 252-262.

Broffitt, J., & Randles, R. H. A power approximation for the chi square goodness-of-fit test: Simple hypothesis case. Journal of the American Statistical Association, 1977, 72, 604-607.

Brownlee, J. Some experiments to test the theory of goodness of fit. Journal of the Royal Statistical Society, 1924, 87, 76-82.

Camilli, G., & Hopkins, K. D. Applicability of chi-square to 2 x 2 contingency tables with small expected cell frequencies.

Psychological Bulletin, 1978, 85, 163-167.

Camp, B. H. Further interpretation of the chi-square test. Journal of the American Statistical Association, 1938, 33, 537-542.

Castellan, J. N. Jr. On the partitioning of contingency tables. Psychological Bulletin, 1965, 64, 330-338.

Chapanis, A. An exact multinomial one-sample test of significance. Psychological Bulletin, 1962, 59, 306-310.

Chapman, D. G., & Meng, R. C. The power of chi-square tests for contingency tables. Journal of the American Statistical Association, 1966, 61, 965-975.

Chapman, J. A. W. A comparison of the X^2, -2 log R, and multinomial probability criteria for significance tests when expected frequencies are small. Journal of the American Statistical Association, 1976, 71, 854-863.

Cochran, W. G. The X^2 distribution for the binomial and Poisson Series with small expectations. Annals of Eugenics, 1936, 2, 207-217.

Cochran, W. G. The X^2 test of goodness of fit. Annals of Mathematical Statistics, 1952, 23, 315-345.

Cochran, W. G. Some methods for strengthening the common X^2 tests. Biometrics, 1954, 10, 417-451.

Conover, W. J. Practical Nonparametric Statistics. New York:

John Wiley & Sons, Inc., 1971.

Conover, W. J. Some reasons for not using the Yates continuity correction on 2 x 2 contingency tables. Journal of the American Statistical Association, 1974a, 69, 374-382.

Conover, W. J. Rejoinder. Journal of the American Statistical Association, 1974b, 69, 382.

Cornfield, J. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast, and cervix. Journal of the National Cancer Institute, 1951, 11, 1269-1275.

Cox, D. R. The continuity correction. Biometrika, 1970, 57, 217-219.

Cramer, H. Mathematical Methods of Statistics. Princeton, New Jersey: Princeton University Press, 1946.

D'Agostino, R. B., & Rosman, B. A normal approximation for testing the equality of two independent chi-square values. Psychometrika, 1971, 36, 251-252.

Dunn, O. J. Multiple comparisons among means. Journal of the American Statistical Association, 1961, 56, 52-64.

Edwards, A. E. On "The use and misuse of the chi-square test": The case of the 2 x 2 contingency table. Psychological Bulletin, 1950, 47, 341-346.

Everitt, B. S. The Analysis of Contingency Tables. London:

Chapman & Hall, 1977.

Fisher, R. A. On the interpretation of X^2 from contingency tables and the calculation of P. Journal of the Royal Statistical Society, 1922, 85, 87-94.

Fisher, R. A. The conditions under which X^2 measures the discrepancy between observation and hypothesis. Journal of the Royal Statistical Society, 1924, 87, 442-450.

Fisher, R. A. Statistical Methods for Research Workers (7th ed.). London: Oliver and Boyd, 1938.

Fisher, R. A. The significance of deviations from expectations in a Poisson series. Biometrics, 1950, 6, 17-24.

Fleiss, J. L. Statistical Methods for Rates and Proportions. New york: Wiley & sons, 1973.

Fry, T. C. The X^2 test of significance. Journal of the American Statistical Association, 1938, 33, 513-525.

Gart, J. J. Approximate confidence limits for the relative risk. Journal of the Royal Statistical Society, Series B, 1962, 24, 454-463.

Gold, R. Z. Tests auxiliary to X tests in a markov chain. Annals of Mathematical Statistics, 1963, 34, 56-74.

Good, I. J., Grover, T. N., & Mitchell, G. J. Exact distributions for X^2 and for the likelihood-ratio statistic for the equiprobable multinomiial distribution. Journal of the

American Statistical Association, 1970, 65, 267-283.

Goodman, L. A. Simultaneous confidence intervals for cross-products ratios in contingency tables. Journal of the Royal Statistical Society, Series B, 1964, 26, 86-102.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. Journal of the American Statistical Association, 1954, 49, 732-764.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. II: Further discussion and references. Journal of the American Statistical Association. 1959, 54, 123-163.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. III: Approximate sampling theory. Journal of the American Statistical Association, 1963, 58, 310,364.

Greenwood, M., & Yule, G. U. The statistics of anti-typhoid and anti-cholera inoculations and the interpretation of such statistics in general. Proceedings of the Royal Society of Medicine, 1915, 8, 113-190.

Grizzle, J. E. Continuity correction in the chi-square test for 2 x 2 tables. American Statistician, 1967, 21 (4), 28-32.

Gumbel, E. J. On the reliability of the classical chi-square test. Annals of Mathematical Statistics, 1943, 14, 253-263.

Harkness, W. L., & Katz, L. Comparison of the power functions for

the test of independence in 2 x 2 contingency tables. The Annals of Mathematical Statistics, 1964, 35, 1115-1127.

Horn, S. D. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. Biometrics, 1977, 33, 237-248.

Irwin, J. O. A note on the subdivision of X^2 into components. Biometrika, 1949, 36, 130-134.

Jeffreys, H. Theory of Probability (3rd ed.). Oxford: Clarendon Press, 1961.

Kastenbaum, M. A. A note on the additive partitioning of chi-square in contingency tables. Biometrics, 1960, 16, 416-422.

Kempthorne, O. The classical problem of inference: Goodness of fit. In J. Neyman (Ed.), Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, California: University of California Press, 1966.

Kendall, M. G. The Advanced Theory of Statistics (Vol. 1, 5th ed.). London: Griffin, 1952.

Kendall, M. G. Rank Correlation Methods (4th ed.). London: Griffin, 1970.

Kendall, M. G., & Stuart, A. The Advanced Theory of Statistics (Vol. 3, 3rd ed.). London: Griffin, 1969.

Kimball, A. W. Short cut formulas for the exact partitioning of X^2 in contingency tables. Biometrics, 1954, 10, 452-458.

Knepp, D. L., & Entwisle, D. R. Testing significance of differences between two chi-squares. Psychometrika, 1969, 34, 331-333.

Kruskal, W. H., & Wallis, W. A. Use of rank in one-criterion variance analysis. Journal of the American Statistical Association. 1952, 47, 401-412.

Lancaster, H. O. The exact partitioning of X^2 and its application to the problem of pooling of small expectations. Biometrika, 1950, 37, 267-270.

Lancaster, H. O. The derivation and partition of X^2 in certain discrete distributions. Biometrika, 1949, 36, 117-129.

Lancaster, H. O. The Chi-Square Distribution. New York: Wiley, 1969.

Lewis, D., & Burke, C. J. The use and misuse of the chi-square test. Psychological Bulletin, 1949, 46, 433-489.

Lewis, D., & Burke, C. J. Further discussion of the use and misuse of the chi-square test. Psychological Bulletin, 1950, 47, 347-355.

Lewontin, R. C., & Felsenstein, J. The robustness of homogeneity tests in 2 x N tables. Biometrics, 1965, 21, 19-33.

Light, R. J., & Margolin, B. H. An analysis of variance for categorical data. Journal of the American Statistical Association, 1971, 66, 534-544.

Mann, H. B., & Wald, A. On the choice of the number of intervals in the application of the chi-square test. Annals of Mathematical Statistics, 1942, 13, 306-317.

Mantel, N. Comment and a suggestion. Journal of the American Statistical Association, 1974, 69, 378-380.

Mantel, N., & Greenhouse, S. W. What is the continuity correction? The American Statistician, 1968, 22 (5), 27-30.

Marascuilo, L. A., & McSweeny, M. Nonparametric and Distribution-Free Methods for the Social Sciences. Monterey, California: Brooks/Cole, 1977.

Maxwell, A. E. Analysing Qualitative Data. London: Methuen & Co., 1961.

Miettinen, O. S. Comment. Journal of the American Statistical Association, 1974, 69, 380-382.

Neyman, J. Contribution to the theory of the X^2 test. In J. Neyman (Ed.), Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, California: University of California Press, 1949.

Neyman, J., & Pearson, E. S. Further notes on the X^2 distribution. Biometrika, 1931, 22, 298-305.

Overall, J. E. Power of chi-square tests for 2 x 2 contingency tables with small expected frequencies. Psychological Bulletin, 1980, 87, 132-135.

Pastore, N. Some comments on "the use and misuse of the chi-square test". Psychological Bulletin, 1950, 47, 338-340.

Patnaik, P. B. The power function of the test for the difference between two proportions in a 2 x 2 table. Biometrika, 1948, 35, 157-175.

Pearson, E. S. The choice of a statistical test illustrated on the interpretation of data classed in a 2 x 2 table. Biometrika, 1947, 34, 139-167.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, July, 1900, pp. 157-175. In E. S. Pearson (Ed.), Karl Pearson's Early Statistical Papers. Cambridge: Cambridge at the University Press, 1947.

Pearson, K. Tables for Statisticians and Biometricians. Cambridge: Cambridge University Press, 1914.

Peters, C. C. The misuse of chi-square: A reply to Lewis and Burke. Psychological Bulletin, 1950, 47, 331-337.

Pirie, W. R., & Hamdan, M. A. Some revised continuity corrections for discrete distributions. Biometrics, 1972, 28, 693-701.

Plackett, R. L. The continuity correction for 2 x 2 tables. Biometrika, 1964, 51, 327-337.

Roscoe, J. T., & Byars, J. A. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. Journal of the American Statistical Association, 1971, 66, 755-759.

Scheffe, H. A method for judging all contrasts in the analysis of variance. Biometrika, 1953, 40, 87-104.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Shaffer, J. P. Testing specific hypotheses in contingency tables: Chi-square partitioning and other methods. Psychological Reports, 1973, 33 (2), 343-348.

Sillitto, G. P. Note on approximation to the power of the 2 x 2 comparitive trial. Biometrika, 1949, 36, 347-352.

Slakter, M. J. Comparative validity of the chi-square and two modified chi-square goodness of fit tests for small but equal expected frequencies. Biometrika, 1966, 53, 619-622.

Slakter, M. J. Accuracy of an approximation to the power of the chi-square goodness of fit test with small but equal expected frequencies. Journal of the American Statistical Association, 1968, 63, 912-924.

Slakter, M. J. A comparison of the Pearson chi-square and Kolmogorov goodness of fit tests with respect to validity. Journal of the American Statistical Association, 1965, 60, 854-858.

Starmer, C. F., Grizzle, J. E., & Sen, P. K. Comment. Journal of the American Statistical Association, 1974, 69, 376-378.

Tate, M. W., & Hyer, L. A. Significance values for an exact multinomial test and accuracy of the chi-square approximation. Final Report No. 8-B-023, Office of Education, Bureau of Research, U.S. D.H.E.W., 1969. (ERIC Reproduction Service No. ED 040 886)

Tate, M. W., & Hyer, L. A. Inaccuracy of the X^2 test of goodness-of-fit when expected frequencies are small. Journal of the American Statistical Association, 1973, 68, 836-841.

West, E. N., & Kempthorne, O. A comparison of the chi and likelihood ratio tests for composite alternatives. Journal of Statistical Computation and Simulation, 1972, 1, 1-33.

Wise, M. E. Multinomial probabilities and the X^2 and χ^2 distributions. Biometrika, 1963, 50, 145-154.

Yarnold, J. K. The minimum expectation in X^2 goodness of fit tests and the accuracy of approximation for the null distribution. Journal of the American Statistical Association, 1970, 65, 864-886.

Yates, F. Contingency tables involving small numbers and the X^2 test. Journal of the Royal Statistical Society Supplement, 1934, 1, 217-235.

Yule, G. U. On the application of the X^2 method of association and contingency tables, with experimental illustrations. Journal

of the Royal Statistical Society, 1922, 85, 95-104.